Extracting Acoustic Features of Singing Voice for Various Applications Related to MIR: A Review

Deepali.Y.Loni[1], Dr. Shaila Subbaraman[2]
[1]Assistant Professor, Textile & Engineering Institute, Department of Electronics, Ichalkaranji, India.
lonidy@dktes.org.in
[2]Professor & HOD, Department of ETC, ADCET, Ashta
shailasubbaraman@yahoo.co.in

***Abstract*— Efficient and intelligent music information retrieval (MIR) is a need of the 21st century. MIR addresses the problem of querying and retrieving certain types of music from large music data set. A singing voice is one of the key elements of music. As most part of music is characterized by the performing singer, analysis of singing voice reveals many characteristics of a song. The unique qualities of a singer's voice make it relatively easy for carrying out numerous tasks in MIR. The singing voice is completely characterized by its acoustic features. Acoustic features like timbre, vibrato, pitch and harmony describe the singing voice in the music and these are discussed in the paper. There are many applications of MIR which considers overall features of music, but the paper presents a review of those applications of MIR concerned directly to singing voice in the music. Also the paper lists the feature extraction methods and identifies the suitable feature appropriate for individual task of MIR.**

***Index Terms*—**Acoustic **Feature Extraction, Classifier, Music Information Retrieval,** Singing voice**,** vocal / non-vocal

I. Introduction

As a major product for entertainment, there is a huge amount of digital musical content produced, broadcasted, distributed and exchanged. This growing demand of amount of music exchange using the internet, and the simultaneous interest of the music industry to find proper means to deal with the new way of distribution, has motivated research activity in the field of MIR [1]. There is a rising demand for music search services. Technologies are demanding for efficient categorization and retrieval of these music collections, so that consumers can be provided with powerful functions for browsing and searching musical content [2].

A singing voice is one of the key elements of music. Singing voice is one of the less studied vocal expressions and analyzing singing voice is an interesting challenge. Singing voice differs from every day speech in its intensity and dynamic range, voiced duration (95% in singing voice whereas 60%in speech), formant structures and pitch. Moreover the singing voice has loud, non-stationary background music signal which makes its analysis relatively more complex than speech [3], [4]. Thus speech processing techniques that were designed for general speech are not always suitable for the singing voice. But major of the earlier work have tried to extend speech processing to the problem of analyzing music signals.

MIR performs the task of classiûcation in which it assigns labels to each song based on genre, mood, artists, etc. Those tasks directly related to song classiûcation analyzing the singing voice are Singer Identification, Singer Verification, Music annotation etc. Other extended applications involve distinguishing between trained and untrained singer, analyse vocal quality, vocal enhancement etc. The paper provides an overview of features and techniques used for the above classiûcation tasks. It provides a summary of different applications based on singing voice and maps the application to its best suitable acoustic feature, the extraction method of that feature and the appropriate classifier. The performance parameters that are essential to evaluate the system are also presented.

II. Basic Framework For Singing Voice Analysis

The singing voice, in addition to being the oldest musical instrument, is also complex from its acoustic standpoint. Processing of the singing voice in a music signal basically involves three major fundamental components: Separation of vocal and non-vocal segments of song, feature extraction from singing part (i.e. vocal segments) which involves analysis of acoustic features and a trained classifier that performs the task of classification and assigns the song to class of the problem.

Separation of vocal and non-vocal segments is an essential component, as most singing voices in popular music are accompanied by musical instruments during vocal passages. Thus the feature vectors extracted from such vocal passages get influenced by the sounds of accompanying instruments. Interference of the instrumental background make an acoustic classifier a poor match to the acoustics of the sung vocal line [5], [6], [7]. Hence, most of the researchers prefer using *a cappella* singing (i.e. with no instrumental background) voice for feature extraction [8]. For accurate analysis of singing voice, it is essential to have a separation of singing voice from the accompanied background sound. Fig. 1 shows the basic steps involved in extracting features from singing voice.

The commonly used techniques for vocal separation are; extraction of feature parameters based on the distribution of energy in different frequency bands, trained hidden Markov models (HMM) as vocal and non-vocal acoustic models, application of melody transcription system to each frame and estimate whether significant melody line is present.

DOI: 03.LSCS.2013.3.520

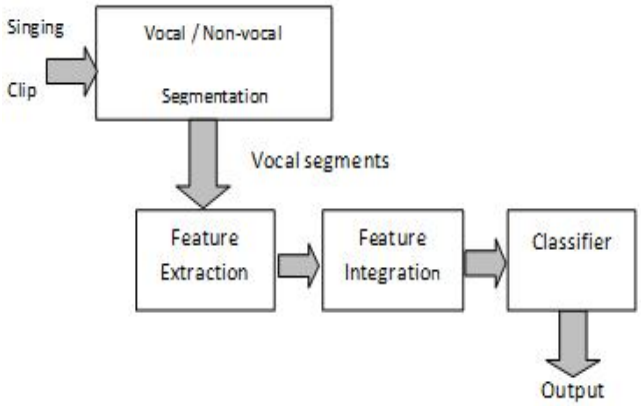

Figure 1. Basic blocks involved in processing of singing voice

After the segmentation of song into vocal/non-vocal regions, features are extracted from the vocal sections. Features are extracted using mathematical transformations like, Wavelet Transform, Mel Frequency Cepstral Coefficients (MFCC), the Linear Prediction Coefficients (LPC) and the Warped Linear Prediction Coefficients.

The feature vectors are then transferred to the classification stage. A classifier is trained using a known set of dataset. When presented with an unknown vocal segment, the classifier assigns the song to the class of problem. The commonly proposed classifiers are Hidden Markov Model (HMM), Neural Networks (NN), Support Vector Machines (SVM) and Gaussian Mixture Model (GMM) classifier.

III. Acoustic Features Of Singing Voice

There are many features that can be extracted from music signal. These features can be categorized into: reference features, content-based features and text-based features. A singing voice can be represented using content-based acoustic features which include timbral texture features, rhythmic content features and pitch content features. Based on these features, singing voice can be analyzed and classified.

A. Pitch Features

It is the perceived fundamental frequency of the sound; which refers to the actual value of the note sung. Pitch refers to the relative lowness or highness that we hear in a sound. The pitch contains features like Pitch Histogram (PH), Pitch Class Profile (PCP) and Harmonic Pitch Class Profile (HPCP) that describe the distribution of pitches. Features such as identifying the highest peak, the amplitude of the highest peak, and the period of the highest peak in the un-folded histogram, selecting the two highest peaks and then compute the distance between the two can be calculated from these pitch content features. Pitch histogram has been used in music genre and mood classiûcation [9], [10] in early years of MIR research. Pitch Class Profile [11] and Harmonic Pitch Class Profile [12] are used in melody analysis and transcription [13], [14], [15].

B. Harmony Features

One of the most discriminative elements to distinguee singing voice from speech is harmonicity. In harmonic sound, the spectral components are the multiples of the lowest (fundamental) frequency. Due to the rapid vibration of the vocal folds, the singing voice is nearly always harmonic [16], and exhibits relatively large amounts of energy at integer multiples of the fundamental frequency in the low or middle frequency regions of the spectrogram. Compared to the singing voices, the instrumental-only sounds have less salient harmonics and spread their energy more widely. The harmonic spectrum is useful in differentiating between low and high pitch singers.

C. Formant Features

Formants are the meaningful frequency components of speech and singing. Formant frequencies are determined by the shape of vocal tract i.e. spectral content of singer's voice. The singer's formant indicates prominent sound energy near 3 kHz and is the result of a clustering of the third, fourth, and fifth formants. This resonance, referred to as the singer's formant, adds a perceptual loudness that allows a singer's voice to be heard over a background accompaniment. Some approaches use the dynamics of F0's (most predominant frequency) trajectory, because a singing voice tends to have temporal variations in its F0 as a consequence of vibrato and such temporal information is expected to express the singer's characteristics. Trained singers often modify the formant structure of their voice in order to add certain desirable characteristics. For example, a lowered second formant results in a "darker" voice–often referred to as "covered"–while a raised second formant produces a "brighter" voice [17]. Trained singers (especially males) often create a resonance in the range of 3000 to 5000 Hz by employing a technique in which the larynx is lowered.

D. Vibrato Features

The pitch of normal speech ranges from 80 to 400 Hz, while that of singing can be from 80 to 1000 Hz. In singing, pitch may be further modulated using a frequency near 4–8 Hz, which results in a phenomenon called vibrato [18]. Vibrato is expressed by vibrating the pitch of the singing voice and is defined as the periodic pitch fluctuation. Singers use vibrato to enhance the expressiveness of their performance. Singing vibrato, however, is an acquired vocal technique and usually requires years to master. Thus vocal vibrato can be seen as a function of the style of singing associated to a particular singer [19]. Some singers have an overly fast vibrato, called tremolo, while others have a wide and slow vibrato, called wobble. Thus vibrato can be considered as an important cue to distinguish between a well-trained singer and a mediocre singer [20]. The vibrato features consist of vibrato rate and vibrato extent. Vibrato rate is the speed of the pitch fluctuation and vibrato extent is the depth of the pitch fluctuation. The rate of vibrato is typically 5-8 Hz, and the modulation depth varies between ±50 and ±150 cents (where 1200 cents = 1 octave) [21].

E. Timbre Features

As a basic element of music, timbre is a term describing

DOI: 03.LSCS.2013.3.520

the quality of a sound [22]. Timbre features are used to differentiate mixtures of sounds that have the same or similar rhythmic and pitch contents. Cleveland [23] states that an individual singer has a characteristic timbre that is a function of the laryngeal source and vocal tract resonances. Timbre is assumed to be invariant with an individual singer and there is a particular range of timbre quality associated to an individual singer.

Many perceptual characteristics of timbre are evident in the spectral content or formant structure of a singer's voice. Thus extraction of timbre features is closely related to spectral analysis of the music signal and requires pre-processing of the signals, which follows some standard steps. Instead of doing song-level signal analysis directly in the first step, a song is usually split into statistically stationary frames, usually by applying a window function at fixed intervals to facilitate subsequent frame- level timbre feature extraction. The application of a window function removes the so-called "edge effects". After framing, spectral analysis techniques such as Fast Fourier Transform (FFT), Short Time Fourier Transform (STFT) and Discrete Wavelet Transform (DWT) are then applied to the windowed signal in each local frame. From the output magnitude spectra of STFT, some timbre features can be defined. Typical timbral features obtained by capturing simple statistics of the spectra include Spectral Centroid (SC), Spectral Rolloff (SR), Spectral Flux (SF), Energy, Zero Crossing and Spectral Bandwidth (SB) [26]. Using DWT a subband analysis can be performed by decomposing the power spectrum into subbands and by applying feature extraction in each subband to extract more powerful features such as MFCC, Octave based Spectral Contrast (OSC), and Daubechies Wavelet Coefficient Histogram (DWCH). Poli [24] measured the timbre quality from spectral envelope of MFCC features to identify singers. In [25], timbre is characterized by the harmonic lines of the harmonic sound.

IV. Choosing The Acoustic Features

The choice of audio features is much dependent on the task to be performed. Timbre features are suitable for genre and instrument classiûcation but not appropriate for comparing the melody similarity of two songs. For mood classiûcation, a large amount of work used rhythm features [27], [28], [29], [30]. While pitch and harmonic features are not quite popular with standard classiûcation systems based on genre, artist, mood, etc., they are the most important feature types for song similarity retrieval and cover song detection at melodic level [31], [32], [33], [34], where timbre features fail to achieve good results. This is corroborated by a recent comparative study on music similarity [35], which showed that timbre features best explain for the instrumentation of the music. Different melodies played by the same instrument would produce more similar timbre features than those corresponding to the same melody with different instrumentation. In general, there is no single set of task-independent features that can consistently outperform the others.

V. Feature Extraction Methods

The separation of vocal part from music accompaniment is potentially very challenging and is the key in providing better solution to singing voice analysis problem. If vital information is lost during this stage, the performance of the following classification stage is inherently crippled and can never measure up to human capability. Typically in singing voice analysis, for each song, spectral features are extracted from frames all over the song and are then clustered to group similar frames together. During feature extraction, the signals are changed into a sequence of feature vectors which are then transferred to the classification stage. If multiple features are available, they can be combined in an effective way to enhance the performance of the system. A good feature extraction approach should have following characteristics, it should be *comprehensive* (represent the music very well), *compact* (require much smaller storage space than the raw acoustic data), and *efficient* (require less computation for extraction) [36].

A. SpectralEnvelope Estimation

In the music field, the signals are usually analyzed in frequency domain. The features of the music signal are always more apparent in frequency domain. The individual characteristics of vocal signals are noticeable in their spectral envelopes [37]. A spectral envelope estimates the vocal tract response. It is a curve which envelopes the magnitude of a short-time spectrum of a signal, linking the peaks or passing close to the maxima of non-sinusoidal spectra.

An effective spectral envelope estimation technique must be capable of handling a wide range of signals with varying characteristics. It is important for a spectral envelope to provide a proper fit to the magnitude spectrum. A certain level of smoothness is desired for a spectral envelope. It should not oscillate erratically, but instead should give a general idea of the distribution of the signal's energy. As the spectral envelope is defined relative to a short segment of the signal (typically between 10 to 50 ms), it should also possess consistency from frame to frame.

Estimation of spectral envelope is the task of deriving spectral envelopes from a given signal. The spectral envelope estimation methods are Linear Predictive Coding, Cepstrum and Discrete Cepstrum.

Linear Predictive Coding (LPC): LPC is an efficient autoregressive class and essentially built up a spectral envelope as the transfer function of an all-pole filter with order 'p' poles. LPC can efficiently indicate the characteristics of harmonic components of the audio signal. Since over 90% of the singing signal is harmonic, LPC is also a good choice for representing the features of a singing voice. The spectral envelope extracted using LPC precisely represents formants of singing voice.

Cepstrum Spectral Envelope: The Cepstrum is a method of speech analysis based on a spectral representation of the signal. After achieving the spectral envelope of the signal, it is possible to analyse the envelope to finds its peak which can provide important data about most relevant formants.

DOI: 03.LSCS.2013.3.520

The Cepstral coefficients derived from LPC analysis has proved to be more robust to noises than the FFT-derived. Cepstral coefficients thus are more appropriate to be used with the singing signal which is mixed with instrumental sounds [37].

B. Mel-Frequency Cepstrum Coefficient(MFCC)

The MFCCs are efficient audio descriptors designed to capture short-term spectral-based features providing spectral energy measurements over short time windows. In order to calculate MFCCs, the signal is first broken into overlapping frames, each approximately 25ms long, a time scale at which the signal is assumed to be stationary. The log-magnitude of the discrete Fourier transform of each window is warped to the Mel frequency scale. A discrete cosine transform (DCT) is then applied and the lower coefficients of the DCT are used to represent a rough shape of the spectrum. By choosing a proper order of the MFCC feature vector, the characteristics of a human voice can be effectively revealed.

The MFCC have been the most successful acoustic features in speech and speaker recognition systems. They have also been successfully used in music signals for artist identification, instrument identification and genre classiûcation.

C. Wavelet Transform

Like Fourier transforms, a wavelet transform is viewed as a tool for dividing signals into different frequency components and then analysing each component with a resolution matched to its scale [38]. Wavelets are designed to give good time resolution at high frequencies and good frequency resolution at low frequencies. After the wavelet decomposition, histogram of each sub-band is constructed. A wavelet coefficients histogram is the histogram of the (rounded) wavelet coefficients obtained by convolving a wavelet filter with an input music signal. Using the wavelet histogram one can calculate statistical features for each sub-band; like the sub-band energy, defined as the mean of the absolute value of coefficients, and the first three moments, i.e., the average, the variance, and the skewness. The histograms of wavelet coefficients, gives a good estimation of the probability distribution over time and thus leads to a good feature representation. Wavelet Transform is used widely in many applications of MIR like classiûcation, similarity, pitch-detection, beat-tracking and indexing problems.

VI. Classifiers

The purpose of classifier learning is to find a mapping from the feature space to the output labels by taking accurate decisions so as to minimize the prediction error. The common choices of classifiers are K-nearest neighbor, support vector machine, and GMM classifier. Various other classifiers have also been used for different music related tasks, including logistic regression, Artificial Neural Networks (ANN), decision trees, Linear Discriminant Analysis (LDA), Nearest Centroid (NC), and Sparse Representation-based Classifier (SRC).

A. K-Nearest Neighbor (K-NN) Classifer

The K-Nearest Neighbors algorithm is the simplest machine learning algorithm, which identifies the object by the majority vote of its neighbors based on distance (usually using a Euclidean distance). Given an input feature vector the algorithm finds k closest feature vectors representing different classes. The disadvantage of K-NN classifier is that its accuracy relies on the selection of an optimum number of neighbors and the most suitable distance measuring method. K-NN has been applied to various music sound analysis problems.

B. Support Vector Machine (SVM)

SVM is the state-of-the-art binary classifier based on the large margin principle and it works well with high- dimensional data [39]. Intuitively, it aims at to construct a hyperplane that divides a data set into n regions, where n is the number of class labels in the data set. These hyperplane simplify to a set of Lagrange multipliers for each training case, and the set of points within the dimensional vectors fed for training that have non-zero Lagrangians are the support vectors. The machine saves these support vectors and applies them to new data in the form of the test set for further on-line classification. Therefore, the SVM has good classiûcation performance since it focuses on the difficult instances.

C. Gaussian Mixture Model (GMM)

The Gaussian mixture model uses multiple weighted Gaussians to attempt to capture the behavior of each class of training data. The use of multiple Gaussians is particularly beneficial when analyzing data that has a distribution not well modeled by a single cluster. It is known that GMMs provide good approximations of arbitrarily shaped densities of a spectrum over a long span of time [40], and hence can reflect the vocal tract configurations of individual singing voice. Hence it is a very flexible model that can adapt to encompass almost any distribution of data. Test points are classified by a maximum likelihood discriminant function, calculated by their distances from the multiple Gaussians of the class distributions [41].To determine the parameters of the Gaussians that best model each class, a well-known technique of Expectation Maximization (EM) is used. EM is an iterative algorithm that converges on parameters that are locally optimal according to the log-likelihood function. It is also useful to perform Principle Components Analysis (PCA) prior to EM. PCA is a multi-dimensional rotation of the data onto the axes of maximal variance. It also has the added benefit of normalizing the data variances, which avoids highly different scaling among the dimensions, which is problematic for EM.

VII. Performance Parameters

Performance parameters are essential to evaluate the system. Some of these, commonly defined in the applications of MIR are:

DOI: 03.LSCS.2013.3.520

A. Accuracy

Every system should compute its accuracy defined as

$$Accu = \frac{Number\ of\ correctly\ identified\ test\ samples}{Total\ number\ of\ samples} \quad (1)$$

B. False Alarm Rate (FAR)

FAR is defined as the number of false alarms divided by the total number of target frames [3], [4].

$$FAR = \frac{Frames falsely\ detected\ as\ target}{Total\ frames\ labelled\ as\ target} \quad (2)$$

C. Miss Detection Rate (MDR)

The miss detection rates are reported as the number of misidentified test samples divided by the number of total test samples [3], [4].

$$MDR = \frac{Frames\ labelled\ as\ target\ but\ undetected}{Total\ target\ frames} \quad (3)$$

VIII. Applications

The ability to capture parameters associated with vocal qualities of singing voice can be applied to a number of tasks in MIR. Some of the applications that have a potential area of research are mentioned below:

- Perform singer identification task to determine who among a group of candidate singers sang a given part of song.
- Evaluation or assessment of the performer's singing ability (performance) in terms of technical accuracy and assigning it a rating score.
- Provide a detail characterization of a particular singing voice to classify it according to skill, style, gender, register, and vocal texture.
- Identify trained and untrained singers by analyzing their acoustic features.
- Perform classical enhancements on the singing voices of untrained singers.
- Perform the singer verification task to decide whether or not a claimed singer performed a given song.
- Convert the music retrieval problem to text retrieval by labeling songs with appropriate tags, substituting songs with text annotations.

A summary of the discussions in the sections III – VI and VIII is put down in Table I.

Table I. lists the different applications based on singing voice and maps the application to its best suitable acoustic feature, the extraction method of that feature and the appropriate classifier.

Conclusions

The vast amount of music accessible to the general public calls for developing tools to effectively and efficiently retrieve and manage the music of interest to the end users. As the singing voice is the basic element of a song that attracts the most attention of listeners; organizing, browsing and classifying music signals based on singing voice is useful for MIR systems.

The complexity of the MIR classification increases with the amount of features used within the classifier. It is therefore crucial to understand the acoustic features and accordingly select only the most relevant features in order to increase the performance of MIR system. The paper has collectively described the acoustic features of singing voice, their extraction methods and the classifiers. Also the paper has put forward some application areas that explore singing voice analysis.

Table I. List of Applications Based on Signing Voice Analysis

MIR Application	Acoustic Features	Extraction Method	Classi -fier	Used in
Singer Identification	Formant features, harmonic features, timbre	MFCC features, LPC	GMM SVM HMM	[1]-[4],[8], [19], [37]
Singer verification	Timbral Features	Cepstrum Coefficient	GMM	[42]
Identify trained/untrained singer	Vibrato, Formant features	LPC, Cepstrum Coefficient	GMM	[20], [43], [44]
Music Annotation	Delta- MFCC x MuVar	MFCC	GMM SVM	[36]
Signal enhancement	MFCC	MFCC	GMM	[20], [45]

References

[1] Andre Holzapfel, Yannis Stylianou, "Singer Identification in Rembetiko Music", Sound and Music Computing Conference (SMC), Lefkada, Greece, 2007.

[2] Tong Zhang, "Automatic Singer Identification", Proceedings of ICME, Baltimore, July 2003.

[3] Tin Lay Nwe and Haizhou Li, "Exploring Vibrato-Motivated Acoustic Features for Singer Identification", IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no.2, pp. 519-530, Feb.2007.

[4] Wei-Ho Tsai, Hsin-Min Wang, "Automatic Singer Recognition of Popular Music Recordings via Estimation and Modeling of Solo Vocal Signals", IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 1, pp. 330-341, January 2006.

[5] Yipeng Li and DeLiang Wang, "Separation of Singing Voice From Music Accompaniment for Monaural Recordings",IEEE Trans. on Audio, Speech, and Language Processing, vol. 15, no. 4, pp.1475-1487, May 2007.

[6] Chao-Ling Hsu, DeLiang Wang, Jyh-Shing Roger Jang, and Ke Hu, "A Tandem Algorithm for Singing Pitch Extraction and Voice Separation From Music Accompaniment", IEEE Trans. on Audio, Speech, and Language Processing,vol.20,no.5,pp.1482-1491, July 2012.

[7] T.L.New, Y.Wang,"Automatic detection of vocal segments in popular songs" Proc. 5th Int. Conf. Music Information Retrieval, Barcelona, Spain, pp.138–145,Oct. 2004.

[8] Wei-Ho Tsai, Hsin-Chieh Lee, "Singer Identification Based on Spoken Data in Voice Characterization", IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 8, pp. 2291-2300, October 2012.

[9] G. Tzanetakis and P. Cook, "Musical genre classiûcation of audio signals," IEEE Trans. Speech Audio Process., vol. 10,

DOI: 03.LSCS.2013.3.520

no. 5, pp. 293–302, 2002.

[10] T. Li and M. Ogihara, "Detecting emotion in music," in Proc. Int. Conf. Music Information Retrieval, 2003.

[11] T. Fujishima, "Realtime chord recognition of musical sound: A system using common lisp music," in Proc. Int. Computer Music Conf., 1999, pp. 464–467.

[12] E. Gomez, "Tonal description of music audio signals," Ph.D. dissertation, Dept. Technol., Universitat Pompeu Fabra, Barcelona, Spain, 2006.

[13] J. Serra, E. Gomez, P. Herrera, andX. Serra, "Chroma binary similarity and local alignment applied to cover song identification," IEEE Trans. Audio, Speech, Lang. Process., vol. 16, no. 6, pp. 1138–1151, 2008.

[14] M. Marolt, "A mid-level melody-based representation for calculating audio similarity,"in Proc. Int. Conf. Music Information Retrieval, 2006.

[15] G. Poliner, D. Ellis, A. Ehmann, E. Gomez, S. Streich, and B. S. Ong, "Melody transcription from music audio: Approaches and evaluation," IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 4, pp. 1247–1256, 2007.

[16] P. R. Cook, "Identification of control parameters in an articulatory vocal tract model, with applications to the synthesis of singing," Ph.D., Stanford Univ., CA, 1990.

[17] Sundberg, J., "The acoustics of the singing voice," Scientific American, vol. 236, pp. 82–91, 1977.

[18] D. Gerhard, "Pitch-based acoustic feature analysis for the discrimination of speech and monophonic singing," J. Canadian Acoust. Assoc., vol.30, no. 3, pp. 152–153, 2002.

[19] Nwe, T.L., Li, H.: Exploring Vibrato-Motivated Acoustic Features for Singer Identification. IEEE Transactions, Audio, Speech and Language Processing 15(2) (2007)

[20] Wei-Ho Tsai, Member, Hsin-Chieh Lee "Automatic Evaluation of Karaoke Singing Based on Pitch, Volume, and Rhythm Features" IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 4, pp. 1233- 1243, May 2012

[21] Sundberg, J., "Human singing voice," in Encyclopedia of Acoustics,pp.1687–1695, John Wiley and Sons, Inc., 1997.

[22] L.Macy, Grove Music Online. [Online]. Available: http:www.oxford-musiconline.com/public/book/omo_gmo.

[23] Cleveland, T.F.: Acoustic Properties of Voice Timbre Types and Their Influence on Voice Classiûcation. Journal of Acoustical Society of America 61, 1622–1629 (1977)

[24] Poli, G.D., Prandoni, P: Sonological Models for Timber Characterization. Journal of New Music Research,170–197

[25] Zhang, T., Kuo, C.C.J.: Content-Based Audio Classiûcation and Retrieval for Data Parsing. Kluwer Academic Publishers, USA (2001)

[26] Zhouyu Fu, Guojun Lu, Kai Ming Ting, and Dengsheng Zhang, "A Survey of Audio-Based Music Classification and Annotation", IEEE Transactions on Multimedia, vol. 13, no. 2, pp. 303-319, April 2011.

[27] Y. Feng, Y. Zhuang, and Y. Pan, "Music retrieval by detecting mood via computational media aesthetics," in Proc. Int. Conf. Web Intelligence, 2003.

[28] D. Yang and W. Lee, "Disambiguating music emotion using software agents," in Proc. Int. Conf. Music Information Retrieval, 2003.

[29] L. Lu, D. Liu, and H.-J. Zhang, "Automatic mood detection and tracking of music audio signals," IEEE Trans. Speech Audio Process., vol.14, pp. 5–18, 2006.

[30] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, "A regression approach to music emotion recognition," IEEE Trans. Audio, Speech, Lang. Process., vol. 16, no. 2, pp. 448–457, 2008.

[31] W. H. Tsai, H. M. Yu, and H. M. Wang, "A query-by-example technique for retrieving cover versions of popular songs with similar melodies," in Proc. Int. Conf. Music Information Retrieval, 2005.

[32] J. Serra, E. Gomez, P. Herrera, and X. Serra, "Chroma binary similarity and local alignment applied to cover song identification," IEEE Trans. Audio, Speech, Lang. Process., vol. 16, no. 6, pp. 1138–1151, 2008.

[33] E. Gomez, "Tonal description of music audio signals," Ph.D. dissertation, Dept. Technol., Universitat Pompeu Fabra, Barcelona, Spain, 2006.

[34] J. P. Bello, "Audio-based cover song retrieval using approximate chord sequences: Testing shifts, gaps, swaps and beats," in Proc. Int. Conf. Music Information Retrieval, 2007.

[35] J. Jensen, M. Christensen, D. Ellis, and S. Jensen, "Quantitative analysis of a common audio similarity measure," IEEE Trans. Audio, Speech, Lang. Process., vol. 17, no. 4, pp. 693–702, 2009.

[36] F. Germain, "The wavelet transform Applications in Music Information Retrieval", McGill University, Canada, December 21, 2009 (citation for B3).

[37] Hiromasa Fujihara, Masataka Goto, Tetsuro Kitahara, Hiroshi Okuno, "A Modeling of Singing Voice Robust to Accompaniment Sounds and Its Application to Singer Identification and Vocal-Timbre-Similarity-Based Music Information Retrieval", IEEE Trans. on Audio, Speech, and Language Processing, vol.18, no.3, pp.638-648, March 2010.

[38] Tao Li, Mitsunori Ogihara, "Toward Intelligent Music Information Retrieval", IEEE Transactions on Multimedia, vol. 8, no. 3, pp. 564-574, June 2006.

[39] B. E. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," in Proc. ACM Conf. Computational Learning Theory, 1992, pp. 144–152.

[40] D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech Audio Process, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[41] R. Duda, P. Hart, and D. Stork, Pattern Classification, 2nd ed. New York: John Wiley & Sons, 2000.

[42] L. Regnier, G. Peeters, "Singer Verification: Singer Model .VS. Song Model", ICASSP 2012, pp. 1-4.

[43] Arun Shenoy, Yuansheng Wu, Ye Wang, "Singing Voice Detection for Karaoke Application", Proc. of SPIE Vol. 5960,2005,pp. 752-762.

[44] Matthew E. Lee, "Acoustic Models for the Analysis and Synthesis of the Singing Voice", Ph.D Thesis, School of Electrical and Computer Engineering, Georgia Institute of Technology, March 2005.

[45] Hideyuki Tachibana, Nobutaka, Ono, Shigeki Sagayama, "Singing Voice Enhancement for Monaural Music Signals Based on Multiple Time-Frequency Analysis", Proceedings of Intersinging 2010 Oct, Tokyo.

DOI: 03.LSCS.2013.3.520